Data mining:

* Data mining is a diverse set of techniques for discovering patterns or knowledge in data.

* Data mining is looking for hidden, valid and potentially useful patterns in huge datasets.

It is also called as knowledge discovery, knowledge extraction, data/pattern analysis, information harvesting etc.

* Techniques of data mining:

(i) Predictive
- (1) classification
- (2) Regression

(ii) Descriptive
- (1) Association
- (2) clustering.

Scanned by CamScanner

Data warehouse:

A data warehouse is a technique for collecting and managing data from varied sources to provide meaningful business insights.

It is a blend of technologies and components which allows the strategic use of data.

* Data warehouse is combining data from multiple sources into one comprehensive and easily manipulated database.

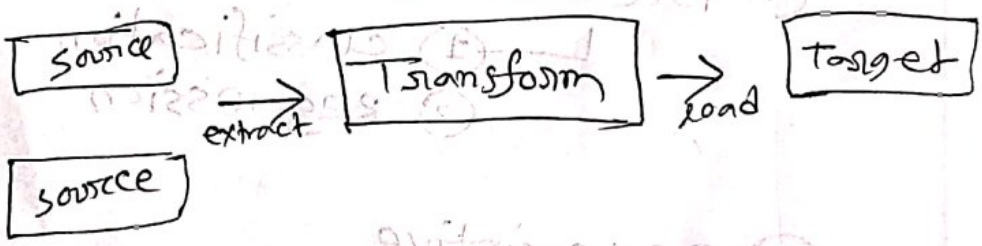

Fig: process of datawarehousing.

Scanned by CamScanner

Normalization:

↳ Normalization is a preprocessing technique used to rescale attributes values to fit in a specific range.

* Techniques of normalization are:

(I) min-max normalization.
(II) Decimal scaling.
(III) Z-score normalization.

(1) min-max normalization:

This is a simple normalization technique in which we fit the data, in a predefined boundary or a predefined interval [C, D].

$$new\_value = \frac{v - min\_value\_of\_dataset}{max\_value - min\_value} * (D - C) + C$$

Scanned by CamScanner

example:

2, 4, 7, 9 ; min = 0, max = 1.

solve

$$new(2) = \frac{2-2}{9-2} * (1-0) + 0$$

$$= \frac{0}{7} * 1$$

$$= 0.$$

$$new(4) = \frac{4-2}{9-2} * (1-0) + 0$$

$$= \frac{2}{7} * 1$$

$$= \frac{2}{7}$$

$$= 0.286.$$

$$new(7) = \frac{7-2}{9-2} * (1-0) + 0$$

$$= \frac{5}{7} * 1$$

$$= 0.714$$

$$new(9) = \frac{9-2}{9-2} * (1-0) + 0$$

$$= \frac{7}{7} * 1$$

$$= 1.$$

Scanned by CamScanner

Z-score:

In this technique, values are normalized based on mean and SD of the data.

$$v' = \frac{v - \bar{x}}{SD}$$

Exmpy

2, 4, 6, 8

$$\bar{x} = \frac{2+4+6+8}{4} = 5$$

$$SD = \sqrt{\frac{(2-5)^2 + (4-5)^2 + (6-5)^2 + (8-5)^2}{4}}$$

$$= 2.236$$

$$\therefore new(2) = \frac{2-5}{2.236} =$$

$$new(4) = \frac{4-5}{2.236} =$$

$$new(6) = \frac{6-5}{2.236} =$$

$$new(8) = \frac{8-5}{2.236} =$$

Scanned by CamScanner

Decimal scaling:

To normalize the data by this technique, we divide each value of the data by the maximum absolute value of data.

$$\therefore v_i' = \frac{v_i}{10^j}$$

example

-10, 201, 301, -401, 501, 601, 701

To normalize the above data,

Step 1: maximum absolute value in given data : 701

Step 2: Divide the given data by 1000. ($\because j=3$).

Result: The normalized data is:

-0.01, 0.201, 0.301, -0.401, 0.501, 0.601

Scanned by CamScanner

Box plot:

A box plot is a graphical representation of statistical data based on the minimum, first quartile, median, third quartile and maximum. The term "box plot" comes from the fact that the graph looks like a rectangle with lines extending from the top and bottom

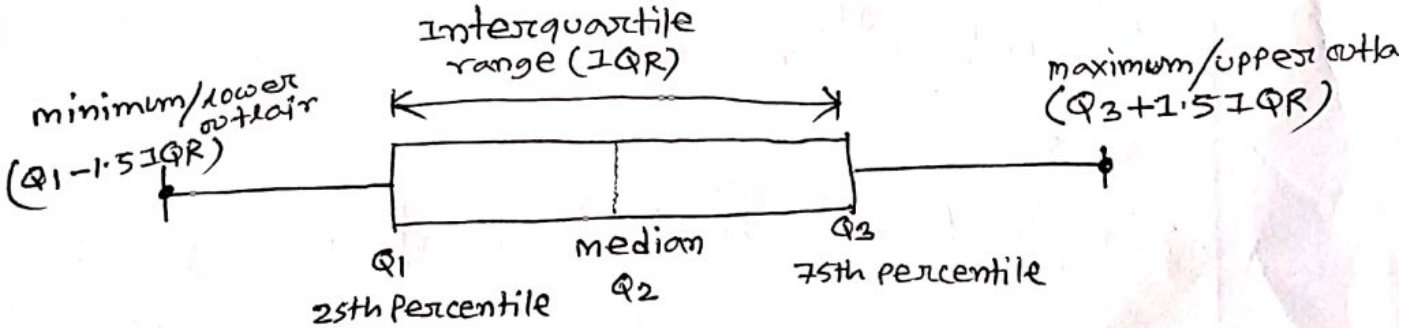

Fig: Different parts of a boxplot.

Scanned by CamScanner

Mehedi
21.10.2019

Classification:

* The separation or ordering of objects (or things) in classes.

* classification is a data mining function that assigns items in a collection to target categories or classes.

The goal of classification is to accurately predict the target class for each case in the data.

For example, a classification model could be used to identify loan applications as low, medium, or high credit risks.

Approaches:

i) Boundary based approaches.
ii) Probability based approaches.
iii) Distance based approaches.
iv) Decision-tree based approaches.
v) Neural-network based approaches.

Missing data:

↳ Missing data means data is not always available.

↳ Missing data values causes problems during both the training phase and the classification process itself. Missing data in the training data must be handled and may produce an inaccurate result.

* Problem arise:

i) Information is not collected.

ii) Attributes may not be applicable to all cases.

* Handle missing data:

↳ Ignore the tuple.

↳ Assume a value for the missing data. This may be determined by using same

method (attribute mean, Bayesian formula or decision tree) to predict what the value could be.

↳ Assume a special value for the missing data. This means that the value of missing data is taken to be a specific value all of its own.

measuring performance:

Let, specific class = c_j
database tuple = t_i

i) True positive (TP):

t_i predicted to be in c_j and is actually in it.

ii) False positive (FP):

t_i predicted to be in c_j but is not actually in it.

iii) True negative (TN):

t_i is not predicted to be in c_j and is not actually in it.

iv) False negative (FN):

t_i is not predicted to be in c_j but is actually in it.

Table 1:

Name	Gender	Height	output
A	F	1.6	S
B	M	2	T
C	F	1.9	M
D	F	1.88	M
E	F	1.7	S
F	M	1.85	M
G	F	1.6	S
H	M	1.7	S
I	M	2.2	T
J	M	2.1	T
K	F	1.8	M
L	M	1.95	M
M	F	1.9	M
N	F	1.8	M
O	F	1.75	M

Bayesian classification:

* Example 4.5:

using the classification results for Table 1, there are four tuples classified as short (S), eight as medium (M) and three as tall (T). To facilitate classification, we divide the height attribute into six ranges:

$(0, 1.6], (1.6, 1.7], (1.7, 1.8], (1.8, 1.9], (1.9, 2.0], (2.0, \infty)$

we classify $t = \{R, M, 1.95\}$.

Solve:

From these training data, we estimate the probabilities:

$$P(S) = \frac{4}{15}$$

$$P(M) = \frac{8}{15}$$

$$P(T) = \frac{3}{15}$$

Counts and subsequent probabilities associated with the attributes are:

Attribute	value	count			probabilities		
		S	M	T	S	M	T
Gender	M	1	2	3	1/4	2/8	3/3
	F	3	6	0	3/4	6/8	0/3
Height	(0, 1.6]	2	0	0	2/4	0/8	0/3
	(1.6, 1.7]	2	0	0	2/4	0/8	0/3
	(1.7, 1.8]	0	3	0	0/4	3/8	0/3
	(1.8, 1.9]	0	4	0	0/4	4/8	0/3
	(1.9, 2.0]	0	1	1	0/4	1/8	1/3
	(2.0, ∞)	0	0	2	0/4	0/8	2/3

we use these values to classify a new tuple.

By using these values and the associated probabilities of gender and height, we obtain the following estimates:

$$P(t/S) = 1/4 * 0/4 = 0$$

$$P(t/M) = \frac{2}{8} * \frac{1}{8} = \frac{1}{32}$$

$$P(t/T) = \frac{3}{3} * \frac{1}{3} = \frac{1}{3}$$

Combining these, we get

likelihood of being S = $P(S) \cdot P(t/S)$

$= \frac{4}{15} * 0 = 0$

likelihood of being M = $P(M) \cdot P(t/M)$

$= \frac{8}{15} * \frac{1}{32} = \frac{1}{60}$

likelihood of being T = $P(T) \cdot P(t/T)$

$= \frac{3}{15} * \frac{1}{3} = \frac{1}{15}$

we estimate $P(t)$ by summing up these individual likelihood values since t will be either S or M or T:

$\therefore P(t) = 0 + \frac{1}{60} + \frac{1}{15} = \frac{1}{12}$

Finally, we obtain the actual probabilities of each event:

$P(S/t) = \frac{P(t/S) \cdot P(S)}{P(t)} = \frac{0 * \frac{4}{15}}{\frac{1}{12}} = 0 = 0\%$

$P(M/t) = \frac{P(t/M) \cdot P(M)}{P(t)} = \frac{\frac{1}{32} * \frac{8}{15}}{\frac{1}{12}} = 0.2 = 20\%$

$P(T/t) = \frac{P(t/T) \cdot P(T)}{P(t)} = \frac{\frac{1}{3} * \frac{12}{15}}{\frac{1}{12}} = 0.799 = 80\%$

so, we classify the new tuple as T because it has the highest probability.

• t = {R, M, 1.95, T}. Ans:

K nearest neighbors (KNN):

* only the K closest entries in the training set are considered further. The new item is then placed in the class that contains the most items from this set of K closest items.

* **Example 4.6:**

Using the sample data from Table 1. we classify the tuple {S, F, 1.6}. Let K=5.

solve:

only the height is used for distance calculations so that both the Euclidean and manhattan distance measures yields the same results.

The distance between A to S = |1.6-1.6| = 0 ... choose(S)
" " B to S = |2-1.6| = 0.4
" " C to S = |1.9-1.6| = 0.3
D to S = |1.88-1.6| = .28
E to S = |1.7-1.6| = 0.1 ... choose(S)
F to S = |1.85-1.6| = .25

The distance between G to S = |1.6 - 1.6| = 0 ..... choose(S)

H to S = |1.7 - 1.6| = 0.1 ..... choose(S)

I to S = |2.2 - 1.6| = 0.6

J to S = |2.1 - 1.6| = 0.5

K to S = |1.8 - 1.6| = 0.2

L to S = |1.95 - 1.6| = 0.35

M to S = |1.9 - 1.6| = 0.3

N to S = |1.8 + 1.6| = 0.2

O to S = |1.75 - 1.6| = 0.15 ..... choose(M)

since K=5, Therefore 5 nearest neighbors to the input tuple are

{A, F, 1.6, S}

{E, F, 1.7, S}

{G, F, 1.6, S}

{H, M, 1.7, S}

{O, F, 1.75, M}

$\therefore S = \frac{4}{5} \times 100 = 80\%$

$M = \frac{1}{5} \times 100 = 20\%$

Of these five items, four are classified as S and one as M. Thus KNN will classify S as S.
(S: name, S: weight)

$\therefore$ {S, F, 1.6, S}.

Ans.